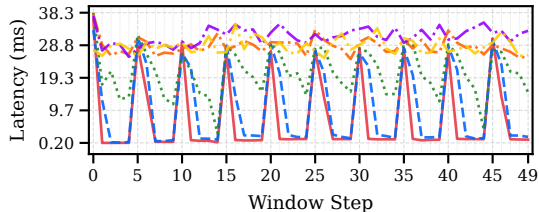


Latency (P50)



— $d_{\text{reduced}} = 16$

... $d_{\text{reduced}} = 128$

- - $d_{\text{reduced}} = 512$

- - $d_{\text{reduced}} = 64$

- . $d_{\text{reduced}} = 256$

- . . Backend Only